Biometrics 60, 249–256
March 2004

A Model to Estimate Risk of Infection with Human Herpesvirus 8 Associated with Transfusion from Cross-Sectional Data

Ruth M. Pfeiffer,[1,*] Sam Mbulaiteye,[2] and Eric Engels[2]

[1]Biostatistics Branch, National Cancer Institute, Bethesda, Maryland 20892-7244, U.S.A.
[2]Viral Epidemiology Branch, National Cancer Institute, Bethesda, Maryland 20892-7244, U.S.A.
[*]*email:* pfeiffer@mail.nih.gov

Summary. In cross-sectional studies of infectious diseases, the data typically consist of: age at the time of study, status (presence or absence) of infection, and a chronology of events possibly associated with the disease. Motivated by a study of how human herpesvirus 8 (HHV-8) is transmitted among children with sickle cell anemia in Uganda, we have developed a flexible parametric approach for combining current-status data with a history of blood transfusions. We model heterogeneity in transfusion-associated risk by a child-specific random effect. We present an extension of the model to accommodate the fact that there is no gold standard for HHV-8 infection and infection status was assessed by a serological assay. The parameters are estimated via maximum likelihood. Finally, we present results from applying various parameterizations of the model to the Ugandan study.

Key words: Age at transfusion data; Current-status data; Mixture models; Multistate models.

1. Introduction

In studies of infectious diseases, epidemiologists frequently wish to estimate and compare the incidence rate between different age groups. Often prospective information is not available, and inference must be based on cross-sectional data consisting of the age and disease status at the time of examination. The time to onset of disease is left-censored for individuals with the disease and right-censored for those without. There is a vast literature on current-status analysis that addresses related issues, much of it motivated by estimating the distribution of the age at HIV infection, time from infection to onset of clinical AIDS, and the per contact probability of virus transmission or infectivity (Shiboski and Jewell, 1992). Approaches to covariate analysis of current-status data were described, e.g., by Rossini and Tsiatis (1996). A general counting process framework for infectious diseases that includes the current-status data setting as a special case was presented by Rhodes, Halloran, and Longini (1996).

This article is motivated by a cross-sectional study of blood-borne transmission of human herpesvirus 8 (HHV-8, also called Kaposi's sarcoma-associated herpesvirus) in 600 Ugandan children afflicted with sickle cell anemia. HHV-8 is the infectious cause of Kaposi's sarcoma, and among children is transmitted through person to person contact, perhaps through saliva. It is uncertain whether blood-borne transmission also occurs, although HHV-8 DNA has been detected at very low levels in peripheral blood. Our primary interest was in capturing the age-specific risk of HHV-8 infection associated with receiving blood transfusions. In children enrolled in our cross-sectional study information on the number of blood transfusions and the ages of transfusions was collected from the mothers using questionnaires. HHV-8 infection status at enrollment was assessed using an assay that measures antibodies against K8.1, an HHV-8 structural glycoprotein. The definition of "infected" was based on predefined cut-off points for the continuous assay optical density readings. Further details about the study and HHV-8 epidemiology can be found in Mbulaiteye et al. (2003a).

By collapsing the transfusion data for each child to consist of the total number of prior transfusions, we could potentially apply existing methods for current-status data (Sun and Kalbfleisch, 1993; Rossini and Tsiatis, 1996), as they do not use information about the times of past events. However, when reliable data are available on the ages of transfusion, efficiency may be gained by incorporating these data into the analysis.

In Section 2 we present a simple parametric model for the distribution of HHV-8 infection, that combines current-status and age-at-transfusion data. This is done by modeling the risk of infection by two components: a discrete component associated with the risk of infection from a blood transfusion at a given age of transfusion, and a continuous component corresponding to infection from other sources, which we refer to as "background infections." This model allows us to estimate the per-transfusion risk of infection, as well as the overall probability of infection in this study. Our model captures individual variation in transfusion risk by including a child-specific random effect in the per-transfusion risk component. This model thus complements existing approaches that would incorporate the number of transfusions as covariate information.

We extend the basic model in Section 3 to accommodate the fact that infection with HHV-8 was assessed using an

antibody assay with imperfect sensitivity and specificity. We use a mixture model for the continuous assay readings that treats the true infection status as a latent variable and thus avoid having to rely on external cut-off values for the classification of "infected." Under this model, the mixing probability, i.e., the probability of true infection in this population, is again composed of two components, namely a transfusion-associated risk component and a "background infection" risk component.

Section 4 provides an application of the model and its extension to data from the Ugandan study, and in Section 5 we discuss our results.

2. The Model and Likelihood

2.1 *Model for Progression of Infection*

Let I_a denote the infection status of a child at his or her current age a, with $I_a = 1$ if the child is infected and 0 otherwise. In what follows we treat age as a continuous variable. The process of infection is characterized by a progressive two-state model with states 0, uninfected, and 1, infected. For a general overview of Markov and multistate models in epidemiology see Commenges (1999). We assume for now that the infection status I of a child is known exactly, an assumption that we relax in the next section to account for the fact that infection status is determined using a serological assay.

The number of transfused units of blood that the child received during his/her lifetime is denoted by n, and $\mathbf{t} = (t_1, \ldots, t_n)$, $t_1 \le t_2 \le \cdots \le t_n$ are the ages at which the transfusions were received. When more than one unit of blood is transfused on the same day, for example the jth and $(j-1)$th unit, then $t_j = t_{j+1}$. The observed data consist of the vectors $(I_a,\ n,\ \mathbf{t})$, if $n > 0$, and $(I_a,\ n = 0)$, otherwise. Complete information on the progression of infection would entail information on the (in our study unobserved) status of infection $I_{t+\delta}$ immediately after each transfusion was received, i.e., $(I_a, n, \mathbf{t}, I_{t_1+\delta}, \ldots, I_{t_n+\delta}) = (I_a, n, \mathbf{t}, \mathbf{I_{t+\delta}})$, where δ denotes a short latency period to detectable infection, thought to be a few weeks for HHV-8. HHV-8 infection, as detected by antibodies, does not clear and thus $I = 1$ is an absorbing state, $P(I_{t_k+\delta} = 1 \,|\, I_{t_{k-1}+\delta} = 1) = 1$ and $P(I_{t_k+\delta} = 0 \,|\, I_{t_{k-1}+\delta} = 1) = 0$. We will comment on this assumption in later sections.

The assumption of an absorbing state in a two-state setting leads to a one-stage Markov model for the progression of infection, that is, the infection status of the person immediately after receiving a transfusion depends only on the infection status following the previous transfusion and the time interval between transfusions, and therefore

$$P(I_a, I_{t_n+\delta}, \ldots, I_{t_1+\delta}) = P(I_a \,|\, I_{t_n+\delta}) P(I_{t_n+\delta} \,|\, I_{t_{n-1}+\delta}) \cdots P(I_{t_1+\delta} \,|\, I_0) P(I_0). \quad (1)$$

Let the random variable T denote the time to transition from $I = 0$ to $I = 1$, or, equivalently, the age at infection. Infection can occur via blood transfusion, at known ages $t_1 \le t_2 \cdots \le t_n$, or from other, perhaps unspecified, causes that we refer to as "background infection." In the absence of information of times or sources of exposure, we postulate a continuous hazard function $\lambda(s)$ for the process of background infection. In the untransfused population, time to infection is governed only by the background rate and thus the hazard function for the background infection at age $s > 0$ can be defined as

$$\lambda(s) = \lim_{\epsilon \to 0^+} \frac{P[s \le T < s + \epsilon \,|\, T \ge s, n = 0]}{\epsilon}.$$

Covariates $X(s)$ could be incorporated, e.g., by specifying a Cox proportional hazards model of the form $\lambda(s) = \lambda_0(s) \exp(X(s)' \beta)$. For an untransfused child, we get

$$\begin{aligned} P(I_a = 1 \,|\, n = 0) &= 1 - P(T > a) \\ &= 1 - P(T > a \,|\, I_0 = 0) P(I_0 = 0) \\ &= 1 - P(I_0 = 0) \exp\left(- \int_0^a \lambda(s)\, ds \right). \end{aligned}$$

While there is conflicting evidence of vertical transmission of HHV-8, in any event the probability of infection at birth is very small (Mantina et al., 2001; Mbulaiteye et al., 2003b). Also, our study does not provide data on infants. Therefore, we assume that $P(I_0 = 0) = 1$.

In the transfused population we have to account for the possibility of infection at the ages of transfusions, $\mathbf{t}$, and specify the hazard function of infection over the infinitesimal interval $[s,\ s + ds)$ by

$$P[s \le T < s + ds \,|\, T \ge s] = \begin{cases} \lambda(s)\, ds + \theta(s), & s = t_i \\ \lambda(s)\, ds, & \text{otherwise,} \end{cases}$$

where the mass points $\theta(t_i)$ correspond to the additional risk of infection associated with a unit blood transfused at age t_i, and again, $\lambda(s)$ is the hazard associated with background infections (for details on hazards with discrete as well as continuous components see for example Kalbfleisch and Prentice [1980], p. 8). As $\lambda(s)ds \ll \theta(s)$, for $s = t_i$ for any i, the above hazard model reduces to

$$P[s \le T < s + ds \,|\, T \ge s] = \begin{cases} \theta(s), & s = t_i \\ \lambda(s)\, ds, & \text{otherwise.} \end{cases}$$

Tied transfusion times are accommodated for by letting $\theta(t_i) = \theta(t_j)$ for $t_i = t_j$. Notice that θ models the probability of infection through any blood transfusion, which depends on the (unknown) prevalence of HHV-8 among blood donors, and not the probability of virus transmission conditional on having received a contaminated transfusion.

We assumed in this setup that the background infection rate is the same in transfused and untransfused children. This is reasonable, as transfusions are related to the severity of anemia, which to our knowledge is not related to risk of viral infection.

If a child who received transfusions is uninfected at the current age a, then there was no infection at any age throughout the interval $[0,\ a]$, and the overall survivor function is

$$\begin{aligned} P(I_a = 0 \,|\, n, \mathbf{t}) &= P(I_a = 0, I_{t_n+\delta} = 0, \ldots, I_{t_1+\delta} = 0) \\ &= P(T > a) \\ &= \prod_{j=1}^{n} (1 - \theta(t_j)) \exp\left(- \int_0^a \lambda(s)\, ds \right) \end{aligned}$$

and

$$P(I_a = 1 \mid n, \mathbf{t}) = 1 - P(I_a = 0 \mid n, \mathbf{t}) = 1 - \prod_{j=1}^{n}(1 - \theta(t_j)) \exp\left(-\int_0^a \lambda(s)\, ds\right), \quad (2)$$

which corresponds to marginalizing over the unobserved intermediate infection states, i.e., all the possible ways that infection might have occurred over time.

2.2 *Parameterization of* $\theta(t)$

The risk of infection per unit blood transfused may vary for different children, i.e., there may be differences in host response to infection due to unobserved factors that influence the immune system such as other infections, e.g., malaria or parasites and malnourishment. To allow for heterogeneity of infection risk among children that is not captured by covariates in the background rate or otherwise in the parametric model, we assume

$$\theta(t \mid Y) = \gamma(t)Y,$$

where Y denotes a child-specific random effect from a distribution with support on (0, 1), and $\gamma(\cdot)$, satisfying $0 \le \gamma(s) \le 1$, is a known function of age at transfusion, that may depend on additional parameters. Unconditionally, the mean risk of infection through transfusion is

$$\theta(t) = \int \theta(t \mid y)\, dF(y) = \gamma(t)E(Y).$$

The probability of being infected for a transfused child is obtained by integrating over the random effect

$$P(I_a = 1 \mid n, \mathbf{t}) = 1 - \exp\left(-\int_0^a \lambda(s)\, ds\right) \times \int \prod_{j=1}^{n}(1 - \theta(t_j \mid y))\, dF(y). \quad (3)$$

A flexible choice of distribution for Y on (0, 1) is the beta distribution, denoted by $B(\alpha, \beta)$, for which the integral in (3) can be written explicitly as follows

$$\begin{aligned}\int \prod_{j=1}^{n}(1 - \gamma(t_j)y)\, dF(y) &= \frac{\Gamma(n+\alpha)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(n+\alpha+\beta)}(-1)^n \prod_{j=1}^{n}\gamma(t_j) \\ &\quad + \frac{\Gamma(n-1+\alpha)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(n-1+\alpha+\beta)}(-1)^{n-1} \\ &\quad \times \sum_{k=1}^{n}\prod_{j\neq k}^{n}\gamma(t_j) + \cdots - \frac{\alpha}{\alpha+\beta}\sum_{k=1}^{n}\gamma(t_k) + 1.\end{aligned}$$

If there is no information on the ages at transfusion, but only on the total number of transfusions, n, we assume $\theta(t \mid Y) = \theta Y$, that is the discrete hazard associated with one unit of transfused blood given the random effect is constant for all ages. The constant θ represent the mean of $\theta(t)$ averaged over all possible transfusion ages $\mathbf{t}$, as $E_{t_1,\ldots,t_n} E_Y \prod_{j=1}^{n}(1 - \theta(t_j, y)) = E_Y \prod_{j=1}^{n}(1 - E_{t_j}\theta(t_j, y))$. In this setting the probability of being uninfected for a transfused individual simplifies to

$$\begin{aligned}P(I_a = 0 \mid n) &= P(T > a \mid n) \\ &= \int (1 - \theta y)^n\, dF(y) \exp\left(-\int_0^a \lambda(s)\, ds\right) \quad (4)\end{aligned}$$

and $P(I_a = 1 \mid n) = 1 - \int (1 - \theta y)^n\, dF(y) \exp(-\int_0^a \lambda(s)\, ds)$, with $\int (1 - \theta y)^n\, dF(y) = \sum_{k=1}^{n}(-\theta)^k n!\Gamma(k+\alpha)\Gamma(\alpha+\beta)/[k!(n-k)!\Gamma(\alpha)\Gamma(k+\alpha+\beta)]$.

Note that from (4), for $\theta(t \mid Y) = \theta$ small,

$$\begin{aligned}P(I_a = 0 \mid n) &= \exp\left(-\int_0^a \lambda(s)\, ds + n\log(1-\theta)\right) \\ &\approx \exp\left(-\int_0^a \lambda(s)\, ds - \theta n\right).\end{aligned}$$

The small, constant transfusion risk θ is thus close to the relative risk estimate based on a log-linear model that treats the number of transfusions as a continuous variable.

Before we present the likelihood for our model, we point out that, if the transfusion risk is modeled using a random effect, the Markov property for I only holds conditionally given Y, with the conditional transition probabilities

$$\begin{aligned}P(I_{t_k+\delta} = 0 \mid I_{t_{k-1}+\delta} = 0, Y) &= P(T > t_k + \delta \mid T > t_{k-1} + \delta, Y) \\ &= \frac{\exp\left(-\int_0^{t_k+\delta} \lambda(s)\, ds\right)\prod_{j=1}^{k}(1 - \theta(t_j \mid Y))}{\exp\left(-\int_0^{t_{k-1}+\delta} \lambda(s)\, ds\right)\prod_{j=1}^{k-1}(1 - \theta(t_j \mid Y))} \\ &= \frac{\exp\left(-\int_0^{t_k+\delta} \lambda(s)\, ds\right)\prod_{j=1}^{k}(1 - Y\gamma(t_j))}{\exp\left(-\int_0^{t_{k-1}+\delta} \lambda(s)\, ds\right)\prod_{j=1}^{k-1}(1 - Y\gamma(t_j))} \\ &= (1 - Y\gamma(t_k)) \exp\left(-\int_{t_{k-1}+\delta}^{t_k+\delta} \lambda(s)\, ds\right).\end{aligned}$$

2.3 *The Likelihood*

We use the subscript i to denote the data $(I_{a_i}, n_i, \mathbf{t}_i)$ or $(I_{a_i}, n_i = 0)$ for the ith child in the sample. The number of untransfused children in the sample is N_0 and N_1 is the number of transfused children. If the infection status I_{a_i} is known for each child in the study, the log-likelihood for the data based on our model is

$$
\begin{aligned}
l_1(\boldsymbol{\theta}, \lambda) &= \log L(\mathbf{I}, \boldsymbol{\theta}, \lambda) \\
&= \sum_{i=1}^{N_0} \left\{ I_{a_i} \log \left(1 - \exp\left(- \int_0^{a_i} \lambda(s)\, ds \right) \right) \right. \\
&\qquad \left. - (1 - I_{a_i}) \int_0^{a_i} \lambda(s)\, ds \right\} \\
&\quad + \sum_{j=1}^{N_1} \left\{ I_{a_j} \log \left(1 - \exp\left(- \int_0^{a_j} \lambda(s)\, ds \right) \right. \right. \\
&\qquad \left. \times \int \prod_{j=1}^{n} (1 - \theta(t_j \,|\, y)) dF(y) \right) \\
&\qquad - (1 - I_{a_j}) \log \left(\int \prod_{j=1}^{n} (1 - \theta(t_j \,|\, y)) dF(y) \right) \\
&\qquad \left. - (1 - I_{a_j}) \int_0^{a_j} \lambda(s)\, ds \right\}. \qquad (5)
\end{aligned}
$$

If there is no information on the ages at transfusion, but only on the total number of transfusions, the data are (I_{a_i}, n_i), and under the assumption that θ is constant over age, the corresponding log likelihood is

$$
\begin{aligned}
l_2(\theta, \lambda) &= \sum_{i=1}^{N_0} \left\{ I_{a_i} \log \left(1 - \exp\left(- \int_0^{a_i} \lambda(s)\, ds \right) \right) \right. \\
&\qquad \left. - (1 - I_{a_i}) \int_0^{a_i} \lambda(s)\, ds \right\} \\
&\quad + \sum_{j=1}^{N_1} \left\{ I_{a_j} \log \left(1 - \exp\left(- \int_0^{a_j} \lambda(s)\, ds \right) \right. \right. \\
&\qquad \left. \times \int (1 - \theta y)^{n_j} dF(y) \right) \\
&\qquad - (1 - I_{a_j}) \log \left(\int (1 - \theta y)^{n_j} dF(y) \right) \\
&\qquad \left. - (1 - I_{a_j}) \int_0^{a_j} \lambda(s)\, ds \right\}. \qquad (6)
\end{aligned}
$$

If the sample contains a mixture of subjects with complete and incomplete information on transfusion times, they can be combined in the likelihood by replacing $\prod_{j=1}^{n} (1 - \theta(t_j \,|\, y))\, dF(y)$ for the children with missing transfusion times in the likelihood (5) by its expectation

$$
E_{\mathbf{t}} \prod_{j=1}^{n} (1 - \theta(t_j \,|\, y))\, dF(y).
$$

If enough data are available, the above expectation can be estimated using the empirical distribution function of $\mathbf{t}$. If the information on $\mathbf{t}$ is sparse, introducing a new parameter, $\theta = E_t \theta(t)$ (in a slight abuse of notation), allows one to directly combine the likelihoods (5) and (6) for children with complete and incomplete transfusion history.

3. A Mixture Model for the Assay Measurements

The true infection status I_a for HHV-8 at age a cannot be observed, as there is no gold standard measure of HHV-8 infection. In our study HHV-8 infection was assessed using an enzyme immunoassay that measures antibodies against K8.1, an HHV-8 structural glycoprotein. The definition of "infected" was based on previous experimental results and a review of the observed distribution of the assay optical density readings, a common approach in epidemiologic studies. Subjects were classified as seronegative (optical density ≤ 0.90), indeterminate (optical density in the range 0.91–1.20), or seropositive (optical density > 1.20), and indeterminate subjects were excluded from the analysis (for details see Mbulaiteye et al., 2003a).

To avoid the somewhat arbitrary definition of cut-off points, to be able to use the "indeterminate" children, and to fully account for the continuous nature of the optical density readings, we model the measurements using a latent class or mixture model that we characterize as follows. Let X denote the optical density measurement for a subject in the study. The observed data for each child thus are $(X_a,\ n,\ \mathbf{t})$ or $(X_a,\ n = 0)$. At age a, each child is in one of two latent or unobserved true infection states, $I_a = 1$ (infected) and $I_a = 0$ (uninfected) with $p = P(I_a = 1)$. The probability density function of X or a transformation of it is modeled as

$$
g(x) = f_0(x)(1 - p) + f_1(x)p,
$$

where f_0 corresponds to the optical density measurements of the uninfected, and f_1 of the infected subpopulation, and the mixing probability p corresponds to the probability of infection in this population. A similar approach for continuous assay measurements was used by Thompson, Smith, and Boyle (1998) to model blood glucose levels, and Pfeiffer, Gail, and Brown (2000) to assess factors that influence infection with *Helicobacter pylori*.

By modeling the state probabilities using $p = P\,(I_a = 1 \,|\, n,\ \mathbf{t})$ or $p = P\ (I_a = 1 \,|\, n = 0)$, from equation (2), we can combine the mixture model for the assay measurements with the model for the progression of infection. After a log transform of the data we chose f_i to be normal with mean μ_i and standard deviation σ_i, for $i = 0, 1$. Given a and the transfusion information n, $\mathbf{t}$, we thus fitted

$$
\begin{aligned}
g(y \,|\, a, n, \mathbf{t}) = {} & f(y; \mu_0, \sigma_0) \cdot (1 - p(a, n, \mathbf{t}; \boldsymbol{\theta}, \lambda)) \\
& + f(y; \mu_1, \sigma_1) \cdot p(a, n, \mathbf{t}; \boldsymbol{\theta}, \lambda), \qquad (7)
\end{aligned}
$$

where $y = \log(x + 0.5)$. We assumed that covariates such as age or transfusion history only influence the probability of infection, but not the f_is. In other applications one could allow the parameters of the distributions to depend on covariates as well.

The likelihood for the mixture model is

$$
l_3(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \log L(\mathbf{Y}, \boldsymbol{\theta}, \lambda, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} \log g(y_i \,|\, a_i). \qquad (8)
$$

We used the EM algorithm to find maximum likelihood estimates for the model parameters ($\boldsymbol{\theta}$, λ, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$) (Titterington, Smith, and Makov, 1985; McLachlan and Basford, 1988). To find confidence intervals for all model estimates we used the

variances from the estimated Fisher information matrix and asymptotic normal theory.

4. Application to the Ugandan HHV-8 Study

4.1 *The Data*

We applied different parameterizations of the model that assumes the infection state to be known (likelihoods [5], [6], and the mixture model [8]) to data that were collected to study the association between HHV-8 infection and transfusion history in Ugandan children with sickle cell anemia. The study includes 600 children attending the Sickle Cell Clinic at Mulago Hospital, Kampala, from November 2001 to April 2002. One child without age information was excluded from the data set. The age range for the remaining 599 children was 1–16 years, as the study was designed to exclude infants, with only one child younger than 1 year. The original design entailed enrolling approximately equal proportions of transfused children and children who had never been transfused. In the final enrollment, never-transfused children comprised 43% (256/599) of the study sample. Interviewers elicited transfusion history from each child's mother and obtained a blood sample that was tested for HHV-8 antibodies using the K8.1 enzyme immunoassay. For roughly half of the transfused children, the mother's recall of transfusion history was validated by reviewing hospital charts. The mothers' reports were highly accurate. Based on preliminary analysis, none of the demographic variables collected in the study were associated with HHV-8 infection status and thus were not included in the models presented in this section (for details see Mbulaiteye et al., 2003a).

4.2 *Analysis with Infection Status Assumed to be Known*

To define "infected" we applied the cut-off points used by Mbulaiteye et al. (2003a), with optical density ≤ 0.90 corresponding to "uninfected," optical density > 1.20 to "infected," and optical density reading in the range 0.90–1.20 to "indeterminate" subjects. The 39 "indeterminate" children as well as 16 children who had missing ages of transfusion were excluded from all analyses in this section, resulting in a data set consisting of 544 children. The number of transfusions, infection status, and ages of the 55 excluded children did not seem different from the rest of the study population.

Table 2 summarizes the results for various parameterizations of the likelihoods (5) and (6).

For Model Ia we fitted the likelihood (6) with a constant background hazard, $\lambda(s) = \lambda$, and a constant transfusion risk θ. Model Ib included the child-specific random effect from a beta distribution with parameters α and β and set $\gamma(s) = 1$. An estimate of the transfusion-specific risk was $\hat{\theta} = \hat{\alpha}/(\hat{\alpha} + \hat{\beta})$ with the 95% CI obtained by applying the delta method. The point estimates for λ and θ and the 95% CIs for both models are very close, with $\hat{\lambda} = 0.028$ and $\hat{\theta} = 0.027(0.004, 0.050)$ for Model Ia, and $\hat{\lambda} = 0.027$ and $\hat{\theta} = 0.029(0.005, 0.053)$ for Model Ib. For both models the transfusion-associated risk was statistically significant. The variance of the beta-distributed random effect was estimated to be 0, indicating little evidence of heterogeneity in transfusion risk among children. This was confirmed by comparing the models using a likelihood ratio test, that resulted in the value 0.12 of the test statistic on one degree of freedom.

Note that in the results presented in Table 1 the age scale for the background infection rate is years, while the time scale for the transfusion risk component is "per transfusion," where the transfusion occurs over a very short time interval, typically a few hours. To make the two estimates comparable with each other one would have to divide λ by 365, which would confirm the assumption that $\lambda \ll \theta$.

Model IIa assumed $\lambda(s) = \lambda$, and let the infection risk associated with transfusion vary by age t of transfusion, that is $\theta(t) = \theta_1$ if $t < 5$, $\theta(t) = \theta_2$ if $5 \leq t < 10$, and $\theta(t) = \theta_3$ if $t \geq 10$. Model IIb included a child-specific random effect from a beta distribution, and let $\gamma(t) = \gamma_1$ if $t < 5$, $\gamma(t) = \gamma_2$ if $5 \leq t < 10$, and $\gamma(t) = \gamma_3$ if $t \geq 10$. Models IIa and IIb yielded very similar estimates for all parameters, and revealed

Table 1

Estimates for various parameterizations of the models for known infection status

	Parameter	MLE (95% CI)	Log likelihood
Model Ia	λ	0.028 (0.021, 0.034)	−262.880
	θ	0.027 (0.004, 0.050)	
Model Ib	λ	0.027 (0.022, 0.032)	−262.831
	α	1.040 (0.140, 1.940)	
	β	34.997 (33.037, 36.957)	
	$\theta = \alpha/(\alpha + \beta)$	0.029 (0.005, 0.053)	
Model IIa	λ	0.025 (−0.006, 0.057)	−260.222
	θ_1, ages [0,5)	0.018 (−0.019, 0.055)	
	θ_2, ages [5,10)	0.075 (0.015, 0.081)	
	θ_3, ages [10,17)	0.000 (−0.126, 0.126)	
Model IIb	λ	0.025 (0.012, 0.038)	−259.847
	γ_1, ages [0,5)	0.175 (−1.137, 1.487)	
	γ_2, ages [5,10)	0.990 (−3.884, 5.864)	
	γ_3, ages [10,17)	0.000 (−2.234, 2.234)	
	α	0.625 (−1.5778, 2.8268)	
	β	5.295 (−41.9190, 52.5087)	
	$\theta_1 = \gamma_1\alpha/(\alpha + \beta)$	0.019 (−0.104, 0.141)	
	$\theta_2 = \gamma_2\alpha/(\alpha + \beta)$	0.104 (−0.418, 0.627)	
	$\theta_3 = \gamma_3\alpha/(\alpha + \beta)$	0.000 (−0.236, 0.236)	

Table 2
Estimates for the mixture model with various parameterizations of the mixing probabilities

	Parameter	MLE (95% CI)	Log likelihood
Model I	λ	0.028 (0.018, 0.037)	−340.307
	θ	0.027 (0.004, 0.050)	
	μ_0	−0.342 (−0.364, −0.319)	
	σ_0	0.099 (0.086, 0.113)	
	μ_1	0.174 (0.157, 0.192)	
	σ_1	0.490 (0.437, 0.544)	
Model II	λ	0.026 (0.017, 0.035)	−333.970
	θ_1, ages [0,5)	0.019 (−0.002, 0.039)	
	θ_2, ages [5,10)	0.078 (0.007, 0.149)	
	θ_3, ages [10,17)	0.000 (−0.439, 0.439)	
	μ_0	−0.348 (−0.366, −0.329)	
	σ_0	0.095 (0.083, 0.107)	
	μ_1	0.164 (0.104, 0.223)	
	σ_1	0.490 (0.450, 0.530)	

heterogeneity in the estimates of $\theta(t)$ for different ages of transfusion. Children who were transfused at 0–5 years of age had a risk of infection of about 0.02, children who were transfused at 5−10 years of age had the highest risk (0.075 for Model IIa, and 0.104 for Model IIb), and both models estimated the risk of infection through transfusion to be zero in the oldest age category. Again, the random-effects model did not fit the data better than the fixed-effects model, based on a likelihood ratio test.

Comparing Model Ia and Model IIa and Models Ib and IIb resulted in the values of 5.32 and 5.97, respectively, on two degrees of freedom for the likelihood ratio test statistic, both significant at the 0.1 level. Letting the transfusion risk vary by age of transfusion thus significantly improved the fit of the models indicating that age is a predictor of risk of infection through transfusion.

Other models without a random effect that let $\lambda(s)$ vary by age did not fit better than models with a constant baseline risk of infection (results not shown). All models estimated λ in the range 2.5–2.8% per year of age, which corresponds to a probability of roughly 0.36 of being infected by age 16 in the untransfused population, in agreement with the prevalences found in other similar populations (Gessain et al., 1999).

To check whether the assumption of the same background rate in transfused and untransfused children is reasonable, we fit Model Ia separately to children without and with transfusions. The estimates of λ were $\hat{\lambda} = 0.028(0.019, 0.360)$ for the untransfused children, and $\hat{\lambda} = 0.023(-0.079, 0.125)$ for the children with at least one transfusion, which justifies using a common parameter for the combined data.

4.3 *Analysis Using the Mixture Model*

As the models incorporating a child-specific effect did not improve the fit of the models, we fitted the mixture models only with fixed transfusion risks. We based the analysis on 583 children, including the 39 children previously classified as "indeterminate," and fit the models to a $\log(x + 0.5)$ transform of the optical density readings. While the mixture model with constant transfusion risk could accommodate the 16 children with missing ages of transfusion, we excluded them to make the analyses comparable.

Several different starting points were used for the EM algorithm, and the algorithm converged for all parameterizations studied in this section.

Table 5 summarizes the results for various parameterizations. For Model I the mixing probability p followed (4), with constant λ and θ. The estimates of $\hat{\lambda} = 0.028$ and $\hat{\theta} = 0.027$ are very close to the model estimates using known infection status, but the confidence intervals are slightly wider. The parameters of the normal mixtures were $\hat{\mu}_0 = -0.342$, $\hat{\sigma}_0 = 0.099$ for the mixture density corresponding to the uninfected children, and $\hat{\mu}_1 = 0.174$ and $\hat{\sigma}_1 = 0.490$ for the density component corresponding to the infected children. The value of the log likelihood was −340.307.

For Model II the parameterization of the mixing probability p incorporated a piecewise constant $\theta(t)$, i.e., $\theta(t) = \theta_1$ if $t < 5$, $\theta(t) = \theta_2$ if $5 \le t < 10$, and $\theta(t) = \theta_3$ if $t \ge 10$. $\lambda(s) = \lambda$ was assumed to be constant.

The estimates were $\hat{\lambda} = 0.026$, $\hat{\theta}_1 = 0.019$, $\hat{\theta}_2 = 0.078$, and $\hat{\theta}_3 = 0$. The parameters of the normal mixtures were $\hat{\mu}_0 = -0.348, \hat{\sigma}_0 = 0.095$, $\hat{\mu}_1 = 0.164$, and $\hat{\sigma}_1 = 0.490$. The estimates for the parameters of the mixing densities were close to those for Model I. The log likelihood was −333.970. The value of the likelihood ratio test was 12.67 on two degrees of freedom, which confirmed the results in the previous section that allowing for heterogeneity in risk of transfusion by ages of transfusion significantly improved the model fit.

For both models the parameter estimates of the mixing probabilities were very close to the estimates based on the fixed cut-off point definition of "infected." The interpretation is slightly different, however, as the mixing probabilities model the latent true infection status, and models (5) and (6) model the event that the assay optical density reading exceeds a certain threshold. The good agreement of the estimates indicates that the chosen cut-offs adequately capture the latent infection state.

4.4 *Goodness-of-Fit*

To test the goodness-of-fit of Models Ia and IIa (assuming known infection status), and the mixture models, we compared the observed numbers of infected children with predicted numbers of infection based on the respective model. The "indeterminate" children were excluded from the

Table 3
Observed and expected number of infected children for the models presented in Tables 1 and 2

No. of transfusions	Model	Age in years <5	5–10	≥10
0	Observed	7/90	15/85	19/63
	Model Ia	7.39	16.46	19.38
	Model IIa	6.99	15.60	18.44
	Mixture model I	7.39	16.46	19.38
	Mixture model II	6.99	15.60	18.44
1–2	Observed	4/57	12/69	21/63
	Model Ia	6.55	14.10	20.27
	Model IIa	5.73	15.50	17.48
	Mixture model I	6.83	15.45	21.98
	Mixture model II	5.73	15.50	17.48
≥3	Observed	8/34	18/47	11/36
	Model Ia	6.21	12.85	12.36
	Model IIa	4.70	14.97	13.41
	Mixture model I	6.63	14.78	14.49
	Mixture model II	4.70	14.97	13.41

analysis to make the models comparable. The first line in each cell of Table 3 shows the observed number of infected subjects in that cell out of the total number of subjects in that cell, the numbers in the lines below are the expected counts of infected individuals based on the respective models. For brevity we do not present the predicted values for Models Ib and IIb, as they were virtually identical to the ones of the corresponding fixed-effects models.

We used as the goodness-of-fit criterion $V_k = \sum (Y_i - n\hat{p}_i)^2 / n\hat{p}_i$ where Y_i is the observed number of infected children in the ith cell, n is the sample size, and $\hat{p}$ is the proportion estimated to be infected under the respective model. The degrees of freedom of the test statistic, k, is defined as the number of nonempty cells minus the number of parameters in the model.

For all models the fit was adequate, although all models underestimated the risk of infection in the 5–10 years age and ≥3 transfusion group. Model IIa and mixture Model II yielded identical estimates of the expected number of infections, which is not surprising as the parameter estimates for the mixing probability were very close for the two models. Model Ia and mixture Model I also resulted in close estimates of the predicted numbers, but the values based on the mixture model were slightly higher in the transfused population. This may reflect the fact that a single parameter does not capture the risk of infection associated with transfusion fully, and the mixture model is slightly more flexible to compensate for that. The values of the test statistic and the associated p-values (given in parentheses) were $V_7 = 0.39(0.99)$ for Model Ia, $V_5 = 0.47(0.99)$ for Model IIa, $V_3 = 0.38(0.94)$ for mixture Model I, and $V_1 = 0.59(0.44)$ for mixture Model II.

5. Discussion

We present a parametric two-state model to estimate the risk of blood-borne transmission of HHV-8 in a cross-sectional sample of children with sickle cell anemia from Uganda. Infection can occur in two different ways, either through blood transfusions at known ages, or through other exposures, collectively referred to as "background infection." By modeling the hazard of infection using a continuous component, corresponding to the background infection rate, and a discrete component at the ages at which transfusions were received, we incorporate the ages of transfusions and thus fully utilize the available data. We compare the model with a simpler model, that assumes a constant hazard associated with transfusion and can be related to relative risk estimates in a log-linear model. We sought to determine whether the risk associated with transfusion varied among children by incorporating a child-specific random effect into the transfusion risk component. This model might detect differences in transfusion risk due to variability in the general levels of immunity in children (which can be modulated by nutritional status, frequency of malaria, or the burden of intestinal parasites), or to variability in the level of circulating virus among blood donors. One has to be careful though, not to overinterpret results on the estimation of the random-effects parameters, as they strongly depend on the model that was chosen for λ. An alternative approach to using a random effect would be to model the background hazard rate nonparametrically, which would cause any heterogeneity among children to be absorbed into the background rate. In an extension, we model the assay measurements using a mixing distribution to avoid using predefined cut-off points for the definition of "infected" in the absence of a gold standard.

The model relies on the assumption that infection as detected by antibodies does not clear. A previous study in homosexual men showed that K8.1 antibody results, once positive, generally remain positive over a period of a decade or more (Biggar et al., 2003). In very young children (less than 1 year of age), HHV-8 antibodies transferred from mothers may clear. However, our study included only one child less than 1 year of age.

Another potential source of bias that would lead to an underestimate of the background infection rate are differences in mortality among uninfected and HHV-8-infected children. While this problem arises in studies of HIV, no differences in mortality by HHV-8 infection status in children have been found (Brayfield et al., 2003).

We fit several parameterizations of the models to data on HHV-8 infection in a cross-sectional sample of children in Uganda. For all models, the overall risk of infection through transfusion was estimated to be around 3% per unit of blood transfused and was found to be statistically significantly different from zero. In our data, we did not detect significant heterogeneity among children in this risk as modeled by the random effect. On the other hand, we did detect heterogeneity in transfusion risk according to the ages at which the transfusions occurred, with lower risk seen in children below 5 years or at least 10 years of age, compared to 5- to 9-year-old children. Notably, the volume of each blood transfusion is calculated directly from the weight of the child. Also, HHV-8 circulates at extremely low levels in infected individuals. Younger children receive smaller volumes of blood than older children, and they may have lower transfusion risk because they are more likely to receive transfusions with small amounts (or even an absence) of virus. For the oldest children, the infection risk might decrease with maturation of the immune system during childhood. These hypothesized explanations need to be examined in additional studies.

Using a mixture model for the assay optical density readings yielded very similar results for risk associated with

transfusion. The fit for all models was significantly improved when transfusion risk was allowed to vary by age of transfusion.

Additional covariates could be incorporated into the model by using a Cox proportional hazards model for the background infection rate, for example. If we assume a constant risk of transfusion, we note that the resulting estimate will be close to the relative risk estimate in a log-linear model that treats the number of transfusions as a continuous variable. Since HHV-8 infection among the children in our study is not a rare event, the relative risk will be different from an odds ratio estimate that is obtained by using the number of transfusions as a continuous variable in a logistic regression model. In the logistic model, the parameter estimates the log odds ratio associated with one additional transfusion. Dunson and Baird (2001) proposed a proportional odds model that estimates the cumulative odds ratio associated with one additional transfusion. Neither of these parameters have the interpretation of the increase in absolute risk associated with transfusion which our method provides.

The study we presented in this article serves as the basis of a larger, longitudinal study that will investigate factors that influence transmission and infection of HHV-8 in Ugandan children.

Acknowledgements

We thank Bob Biggar, Nilanjan Chatterjee, Mitchell Gail, and Steve Mark for helpful comments and discussions.

Résumé

Dans les études transversales de maladies infectieuses, les données consistent typiquement en : l'âge au moment de l'étude, le statut de l'infection (présence ou absence) et une chronologie d'événements pouvant être associés à la maladie. Motivés par une étude sur la transmission de l'herpes virus humain 8 (HHV-8) parmi les enfants ayant une drépanocytose en Ouganda, nous avons développé une approche paramétrique flexible pour combiner les données du statut et l'histoire des transfusions sanguines. Nous modélisons l'hétérogénéité du risque associé à la transfusion par un effet aléatoire spécifique à l'enfant. Nous présentons une extension du modèle pour s'accommoder du fait qu'il n'existe pas une référence standard pour l'infection HHV-8 et que le statut de l'infection a été mesuré par un test sérologique. Les paramètres sont estimés par le maximum de vraisemblance. Enfin, nous présentons les résultats de diverses paramétrisations du modèle appliquées à l'étude Ougandaise.

References

Biggar, R. J., Engels, E. A., Whitby, D., Kedes, D. H., and Goedert, J. J. (2003). Antibody reactivity to latent and lytic antigens to human herpesvirus-8 in longitudinally followed homosexual men. *Journal of Infectious Diseases* **187,** 12–18.

Brayfield, B. P., Phiri, S., Kankasa, C., Muyanga, J., Mantina, H., Kwenda, G., West, J. T., Bhat, G., Marx, D. B., Klaskala, W., Mitchell, C. D., and Wood, C. (2003). Postnatal human herpesvirus 8 and human immunodeficiency virus type 1 infection in mothers and infants from Zambia. *Journal of Infectious Diseases* **187,** 559–5681.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis* **5,** 315–327.

Dunson, D. B. and Baird, D. D. (2001). A flexible parametric model for combining current status and age at first diagnosis data. *Biometrics* **57,** 396–403.

Gessain, A., Mauclere, P., van Beveren, M., Plancoulaine, S., Ayouba, A., Essame-Oyono, J. L., Martin, P. M. V., and de The, G. (1999). Human herpesvirus 8 primary infection occurs during childhood in Cameroon, Central Africa. *International Journal of Cancer* **81,** 189–192.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley, p. 163–178.

Mantina, H., Kankasa, C., Klaskala, W., Brayfield, B., Campbell, J., Du, Q. J., Bhat, G., Kasolo, F., Mitchell, C., and Wood, C. (2001). Vertical transmission of Kaposi's sarcoma-associated herpesvirus. *International Journal of Cancer* **94,** 749–752.

Mbulaiteye, S. M., Biggar, R. J., Bakaki, P. M., Pfeiffer, R., Whitby, D., Owor, A. M., Katongole-Mbidde, E., Goedert, J. J., Ndugwa, C. M., and Engels, E. A. (2003a). Blood-borne transmission of human herpesvirus 8 in children with sickle cell disease in Uganda. *Journal of the National Cancer Institute* **95,** 1330–1335.

Mbulaiteye, S. M., Pfeiffer, R. M., Whitby, D., Brubaker, G. R., Shao, J., and Biggar, R. J. (2003b). Human herpesvirus 8 infection within families in rural Tanzania. *Journal of Infectious Diseases* **187,** 1780–1785.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

Pfeiffer, R., Gail, M. H., and Brown, L. (2000). A mixture model for the distribution of IgG antibodies to *Helicobacter pylori*: Application to studying factors that affect prevalence. *Journal of Epidemiology and Biostatistics* **5,** 267–275.

Rhodes, P. H., Halloran, M. E., and Longini, I. M. (1996). Counting process models for infectious disease data: Distinguishing exposure to infection from susceptibility. *Journal of the Royal Statistical Society Series B* **58,** 751–762.

Rossini, A. J. and Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* **91,** 713–721.

Shiboski, S. C. and Jewell, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association* **87,** 360–372.

Sun, J. G. and Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. *Journal of the American Statistical Association* **88,** 1449–1454.

Thompson, T. J., Smith, P. J., and Boyle, J. P. (1998). Finite mixture models with concomitant information: Assessing diagnostic criteria for diabetes. *Applied Statistics* **47,** 393–404.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Received January 2003. *Revised July* 2003.
Accepted September 2003.